ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Foundations of Data Science
Fall 2025

Assignment date: Saturday, January 31, 2026, 9:15
Due date: Saturday, January 31, 2026, 12:15

Final Exam – CE 1 4

This exam is open book. No electronic devices of any kind are allowed. There are 4 problems. Choose the ones you find easiest and collect as many points as possible. Good luck!

Name: _____

Problem 1	/ 10
Problem 2	/ 10
Problem 3	/ 10
Problem 4	/ 10
Total	/40

Problem 1 (A perspective on PCA – 10 pts)**.** Let $\mathbf{X} \in \mathbb{R}^d$ be a zero-mean random vector with covariance matrix

$$\mathbf{\Sigma} = \mathbb{E}[\mathbf{X}\mathbf{X}^\top].$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ denote the eigenvalues of $\mathbf{\Sigma}$.

(i) (2 pts) Let $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 = 1$. Show that

$$\mathrm{Var}(\mathbf{w}^\top \mathbf{X}) = \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}.$$

(ii) (4 pts) Show that the unit vector $\mathbf{w}$ maximizing $\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}$ is the eigenvector corresponding to the largest eigenvalue λ_1.

(iii) (3 pts) Let $\mathbf{u}_1$ be the eigenvector associated with λ_1, and define the rank-1 reconstruction

$$\hat{\mathbf{X}} = \mathbf{u}_1 \mathbf{u}_1^\top \mathbf{X}.$$

Compute the rank-1 reconstruction error

$$\mathbb{E}\left[\|\mathbf{X} - \hat{\mathbf{X}}\|_2^2\right].$$

(iv) (1 pt) Define the explained variance ratio of the first principal component as

$$\mathrm{EVR}(1) = \frac{\lambda_1}{\sum_{i=1}^d \lambda_i}.$$

Discuss and interpret EVR(1) in words. Then, give a practical criterion for deciding whether a rank-1 PCA approximation is sufficient.

Solution 1.

(a)

Since $\mathbb{E}[\mathbf{X}] = 0$,

$$\mathrm{Var}(\langle \mathbf{w}, \mathbf{X} \rangle) = \mathbb{E}[(\mathbf{w}^\top \mathbf{X})^2] = \mathbb{E}[\mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w}] = \mathbf{w}^\top \mathbb{E}[\mathbf{X}\mathbf{X}^\top] w = \mathbf{w}^\top \Sigma \mathbf{w}.$$

(b)

Option 1 (Recommended): Spectral Decomposition Argument

Let $\{\mathbf{u}_i\}_{i=1}^d$ be an orthonormal eigenbasis of $\mathbf{\Sigma}$, with $\mathbf{\Sigma} = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Write

$$\mathbf{w} = \sum_{i=1}^d \alpha_i \mathbf{u}_i,$$

then $\|\mathbf{w}\|_2^2 = \sum_{i=1}^d \alpha_i^2 = 1$.

Then

$$\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} = \left(\sum_{i=1}^d \alpha_i \mathbf{u}_i \right)^\top \left(\sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \right) \left(\sum_{i=1}^d \alpha_i u_i \right) = \sum_{i=1}^d \lambda_i \alpha_i^2 \leq \lambda_1 \sum_{i=1}^d \alpha_i^2 = \lambda_1.$$

Equality holds if and only if $\alpha_i = 0$ for all i such that $\lambda_i < \lambda_1$. In particular, choosing $\mathbf{w} = \mathbf{u}_1$ yields

$$\max_{\|\mathbf{w}\|_2 = 1} \mathbf{w}^\top \Sigma \mathbf{w} = \lambda_1.$$

Option 2: Lagrange Multiplier Argument

We maximize $\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}$ subject to the constraint $\|\mathbf{w}\|_2^2 = 1$. Consider the Lagrangian

$$\mathcal{L}(\mathbf{w}, \mu) = \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} - \mu(\mathbf{w}^\top \mathbf{w} - 1).$$

Setting the gradient with respect to $\mathbf{w}$ to zero gives

$$2\mathbf{\Sigma} \mathbf{w} - 2\mu \mathbf{w} = 0,$$

which implies

$$\mathbf{\Sigma} \mathbf{w} = \mu \mathbf{w}.$$

Hence any stationary point must be an eigenvector of $\mathbf{\Sigma}$ corresponding to eigenvalue λ, and the objective value at such a point is

$$\mathbf{w}^\top \mathbf{\Sigma} w = \mu \|\mathbf{w}\|_2^2 = \mu.$$

Therefore the maximum is attained by choosing $\mathbf{w}$ as an eigenvector associated with the largest eigenvalue λ_1, and the maximum value equals λ_1.

(c)

Using the eigendecomposition $\mathbf{\Sigma} = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ and the orthonormal basis $\{\mathbf{u}_i\}$,

$$\mathbf{X} = \sum_{i=1}^d \langle \mathbf{u}_i, \mathbf{X} \rangle \mathbf{u}_i.$$

The reconstruction keeps only the first component:

$$\hat{\mathbf{X}} = \langle \mathbf{u}_1, \mathbf{X} \rangle \mathbf{u}_1.$$

Hence,

$$\mathbf{X} - \hat{\mathbf{X}} = \sum_{i=2}^{d} \langle \mathbf{u}_i, \mathbf{X} \rangle \mathbf{u}_i, \quad \|\mathbf{X} - \hat{\mathbf{X}}\|_2^2 = \sum_{i=2}^{d} \langle \mathbf{u}_i, \mathbf{X} \rangle^2.$$

Taking expectation,

$$\mathbb{E}[\|\mathbf{X} - \hat{\mathbf{X}}\|_2^2] = \sum_{i=2}^{d} \mathbb{E}[\langle \mathbf{u}_i, \mathbf{X} \rangle^2] = \sum_{i=2}^{d} \mathbf{u}_i^\top \boldsymbol{\Sigma} \mathbf{u}_i = \sum_{i=2}^{d} \lambda_i.$$

(d)

1. EVR(1) is the fraction of the total variance of the data explained by the first principal component.

2. In practice, a rank-1 approximation is considered sufficient if EVR(1) exceeds a chosen threshold (e.g. 90% or 95%), or if adding further components yields diminishing returns ("elbow" criterion).

Problem 2 (The Log-Likelihood Ratio – 10 pts)**.** Given samples $X_1, X_2, \ldots, X_n$, we studied the problem of deciding whether the data consists of i.i.d. samples from distribution P_0 or from distribution P_1. We saw that the key is the log-likelihood ratio

$$\Lambda_n(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{P_1(X_i)}{P_0(X_i)}. \tag{1}$$

Now, let us assume that indeed, $X_1, X_2, \ldots, X_n$, are i.i.d. samples from P_1. In that case, we showed in class that $\mathbb{E}_{P_1}[\Lambda_n(X_1, \ldots, X_n)] = D(P_1 \| P_0)$. Give a good *upper* bound on the probability

$$\mathbb{P}_{P_1} \left\{ |\Lambda_n(X_1, \ldots, X_n) - D(P_1 \| P_0)| \geq \eta \right\}. \tag{2}$$

The better your bound, the more points. As always, full step-by-step justifications are required for full credit.

Hint: You may start by assuming that X_i are binary. In that case, P_1 is a Bernoulli(α_1) distribution and P_0 is a Bernoulli(α_0) distribution.

Solution 2. We present two different approaches to tackle this problem.

1. **Chernoff/Hoeffding approach:** Given the contents of our class, this is what we expected you would do. Namely, leveraging what we saw in class, we study the random variables

   $$Z_i = \log \frac{P_1(X_i)}{P_0(X_i)}, \tag{3}$$

   where X_i is distributed according to P_1.

   The **key observation is that the random variables Z_i are i.i.d.**, that is, they are independent and identically distributed. This is because the X_i are i.i.d., and each Z_i is a deterministic function ("remapping") of the corresponding X_i. Moreover, as we saw in class (and as is given in the problem statement), the mean of Z_i (when X_i is distributed according to P_1) is

   $$\mu := \mathbb{E}_{P_1}[Z_i] = D(P_1 \| P_0). \tag{4}$$

   With this notation, we can write the quantity that we need to bound as follows:

   $$\mathbb{P}_{P_1} \left\{ |\Lambda_n(X_1, \ldots, X_n) - D(P_1 \| P_0)| \geq \eta \right\} = \mathbb{P}_{P_1} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mu \right| \geq \eta \right\}. \tag{5}$$

This is exactly of the form that we have studied again and again in class (including, for example, in the context of multi-armed bandits). Specifically, just like in class, we can write

$$\mathbb{P}_{P_1}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq \eta\right) = \mathbb{P}_{P_1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu \geq \eta\right) + \mathbb{P}_{P_1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu \leq -\eta\right). \tag{6}$$

To use the tools we learned in class, we would now like to assert that the Z_i are subgaussian and find the corresponding variance proxy σ^2. Clearly, if we can establish this, then we immediately, without any further thinking, have the bound

$$\mathbb{P}_{P_1}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq \eta\right\} \leq 2e^{-\frac{n\eta^2}{2\sigma^2}}, \tag{7}$$

from Lemma 2.6 in the lecture notes (Hoeffding's Bound).

But is there reason to believe that Z_i are indeed subgaussian? — This is where the hint comes in. Namely, starting with the simple binary setting, we realize that the random variable Z_i only has two different values, namely,

$$Z_i = \begin{cases} \log\frac{1-\alpha_1}{1-\alpha_0}, & \text{with probability } 1-\alpha_1 \\ \log\frac{\alpha_1}{\alpha_0}, & \text{with probability } \alpha_1 \end{cases} \tag{8}$$

Let us assume that $0 < \alpha_0 < 1$ **and** $0 < \alpha_1 < 1$. In this case, $Z_i \in [a, b]$ is a bounded random variable, and the length of the interval is

$$b - a = \left|\log\frac{1-\alpha_1}{1-\alpha_0} - \log\frac{\alpha_1}{\alpha_0}\right|, \tag{9}$$

which is finite. (We discuss the more general case below.) By Lemma 2.5 from the Lecture Notes, we can thus assert that Z_i are subgaussian with variance proxy $\left(\log\frac{(1-\alpha_1)\alpha_0}{(1-\alpha_0)\alpha_1}\right)^2/4$. Hence, we can give the following bound:

$$\mathbb{P}_{P_1}\left(|\Lambda_n(X_1, \ldots, X_n) - D(P_1\|P_0)| \geq \eta\right) \leq 2e^{-\frac{n\eta^2}{2\sigma^2}}, \tag{10}$$

where $\sigma^2 = \left(\log\frac{(1-\alpha_1)\alpha_0}{(1-\alpha_0)\alpha_1}\right)^2/4$. **This is the bound we expected you to find.**

How to generalize this to the case where the X_i are not just binary? Let us continue with X_i that are discrete, supported on an alphabet of size k. Then, again, we can assert that Z_i only takes k different values. **Assuming all of these values are finite,** we again directly have that Z_i is a bounded random variable. And the interval is simply given by

$$\left[\min_x \log\frac{P_1(x)}{P_0(x)}, \max_x \log\frac{P_1(x)}{P_0(x)}\right]. \tag{11}$$

If you want to know more (but nothing of the sort was expected!): To beautify notation, let us define

$$D_\infty(P_1\|P_0) = \max_x \log \frac{P_1(x)}{P_0(x)}. \tag{12}$$

(This is a standard definition in the literature, called *Rényi divergence of order* ∞.) With this, we can express the interval as

$$[-D_\infty(P_0\|P_1), D_\infty(P_1\|P_0)]\,, \tag{13}$$

and the length of the interval can be written as $D_\infty(P_1\|P_0) + D_\infty(P_0\|P_1)$. Note the pleasing fact that this formula is symmetric in P_0 and P_1. That is, the variance proxy is given by $\sigma^2 = (D_\infty(P_1\|P_0) + D_\infty(P_0\|P_1))^2/4$. And with this, just for kicks, we can write the full bound as

$$\mathbb{P}_{P_1}\left(|\Lambda_n(X_1,\ldots,X_n) - D(P_1\|P_0)| \geq \eta\right) \leq 2e^{-\frac{2n\eta^2}{(D_\infty(P_1\|P_0)+D_\infty(P_0\|P_1))^2}}. \tag{14}$$

An interesting follow-up discussion concerns the case when there exists an x such that $\log \frac{P_1(x)}{P_0(x)}$ is **infinite**. Evidently, we can exclude all values of x for which $P_1(x) = 0$. Under P_1, these values do not even show up. So the issue is only that there is an x for which $P_0(x) = 0$ (but $P_1(x) > 0$). For such value of x, we have $\log \frac{P_1(x)}{P_0(x)} = \infty$. The simple key observation now is that in this case, the Kullback-Leibler divergence $D(P_1\|P_0)$ (and thus, the mean μ) is also infinite. Hence, this case is uninteresting in the sense that we cannot give any bound of the type that we are looking for (that is better than the trivial upper bound of 1).

Finally, you may be interested in the case of general real-valued random variables. Extrapolating from above, we can see that if our distributions $P_0(x)$ and $P_1(x)$ are such that $\min_x \log \frac{P_1(x)}{P_0(x)}$ and $\max_x \log \frac{P_1(x)}{P_0(x)}$ are both finite, then we are done. Namely, we again get the variance proxy given by $\sigma^2 = (D_\infty(P_1\|P_0) + D_\infty(P_0\|P_1))^2/4$.

2. **Chebyshev approach:** Another approach is to directly plug into Equation (2.4) from the lecture notes, which is the Chebyshev inequality. Namely,

$$\mathbb{P}_{P_1}\left\{|\Lambda_n(X_1,\ldots,X_n) - D(P_1\|P_0)| \geq \eta\right\} \leq \frac{\mathrm{Var}(\Lambda_n(X_1,\ldots,X_n))}{\eta^2}, \tag{15}$$

where we now have to find (or at least bound) the variance of $\Lambda_n(X_1,\ldots,X_n)$ when the X_i are i.i.d. according to P_1. To do this, we note

$$\mathrm{Var}(\Lambda_n(X_1,\ldots,X_n)) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \log\frac{P_1(X_i)}{P_0(X_i)}\right) \tag{16}$$

$$= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left(\log\frac{P_1(X_i)}{P_0(X_i)}\right), \tag{17}$$

where the last step is an argument that we have used several times during the semester: The random variables $Z_i := \log \frac{P_1(X_i)}{P_0(X_i)}$ are independent (and identically distributed) random variables. Therefore, the variance of the sum is equal to the sum of the variances (easy to prove — do it again if you are unsure!). Finally, because the Z_i are also identically distributed (and hence, all the variance terms in the sum are equal), we can write our bound as

$$\mathbb{P}_{P_1}\left\{|\Lambda_n(X_1,\ldots,X_n) - D(P_1\|P_0)| \geq \eta\right\} \leq \frac{\mathrm{Var}\left(\log \frac{P_1(X)}{P_0(X)}\right)}{n\eta^2}, \tag{18}$$

where the dummy random variable X is distributed according to P_1. This is already a nice result: Since the variance term does not depend on n, the bound goes to zero as n becomes large and is thus an interesting and non-trivial bound. (Although, as we emphasized in class, this bound is only inversely proportional to n, by contrast to the Chernoff/Hoeffding approach, which leads to an exponential decay as a function of n.) This is true as long as the variance term is finite.

So, the next question is: What can we say about the variance term? This is where the hint comes in: Let us take P_0 and P_1 to be Bernoulli distributions, as suggested. Then, (as above) we know that

$$Z := \log \frac{P_1(X)}{P_0(X)} = \begin{cases} \log \frac{1-\alpha_1}{1-\alpha_0}, & \text{with probability } 1-\alpha_1 \\ \log \frac{\alpha_1}{\alpha_0}, & \text{with probability } \alpha_1 \end{cases} \tag{19}$$

It is straightforward (but perhaps a bit tedious) to express the variance of this random variable. For example,

$$\mathrm{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \tag{20}$$

$$= (1-\alpha_1)\left(\log \frac{1-\alpha_1}{1-\alpha_0}\right)^2 + \alpha_1\left(\log \frac{\alpha_1}{\alpha_0}\right)^2 - \left(D(P_1\|P_0)\right)^2, \tag{21}$$

which (as long as $0 < \alpha_0 < 1$) is a well-defined finite number.

Alternatively, you can also just bound the variance. A very simple bound for any random variable W is $\mathrm{Var}(W) \leq \max_w |w|^2$, where the maximum can be limited to all values w for which $p_W(w) > 0$. (Of course, this bound is only interesting if this maximum value is finite. Otherwise, more careful work is needed.) For our case, we can write this as

$$\mathrm{Var}\left(\log \frac{P_1(X)}{P_0(X)}\right) \leq \left(\max\left\{-\min_x \log \frac{P_1(x)}{P_0(x)}, \max_x \log \frac{P_1(x)}{P_0(x)}\right\}\right)^2 \tag{22}$$

$$= \left(\max\{D_\infty(P_0\|P_1), D_\infty(P_1\|P_0)\}\right)^2 \tag{23}$$

$$\leq \left(D_\infty(P_0\|P_1) + D_\infty(P_1\|P_0)\right)^2, \tag{24}$$

where we have only added the last, loose bounding step to connect to the Chernoff/Hoeffding approach. Namely, via Chebyshev, we can thus establish the bound

$$\mathbb{P}_{P_1}\left\{|\Lambda_n(X_1,\ldots,X_n) - D(P_1\|P_0)| \geq \eta\right\} \leq \frac{\left(D_\infty(P_1\|P_0) + D_\infty(P_0\|P_1)\right)^2}{n\eta^2}, \tag{25}$$

while Chernoff-Hoeffding permitted to have the bound

$$\mathbb{P}_{P_1}\left(|\Lambda_n(X_1,\ldots,X_n) - D(P_1\|P_0)| \geq \eta\right) \leq 2e^{-\frac{2n\eta^2}{\left(D_\infty(P_1\|P_0)+D_\infty(P_0\|P_1)\right)^2}}. \tag{26}$$

Problem 3 (Projection Theorem – 10 pts)**.** Consider a Hilbert space H spanned by the orthonormal basis $\{\mathbf{z}_i\}_{i\in\mathbb{Z}_+}$. Let $G \subseteq H$ be a (Hilbert) subspace of H, spanned by the first N basis vectors, that is, $G = \text{span}\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$.

(i) (3 pts) For a fixed vector $\mathbf{d} \in H$, give an expression for $\min_{\hat{\mathbf{d}}\in G} \|\mathbf{d} - \hat{\mathbf{d}}\|^2$ in terms of $\mathbf{d}$ and $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$:

$$\min_{\hat{\mathbf{d}}\in G} \|\mathbf{d} - \hat{\mathbf{d}}\|^2 = \ldots \tag{27}$$

The simpler your expression, the more points you get. *Hint: Check your lecture notes. Recall that we are given an orthonormal basis of G!*

(ii) (1 pt) Now let H be the space of all zero-mean finite variance (real-valued) random variables. Here, we prefer to denote the abstract vectors $\mathbf{d}$ and $\mathbf{z}_i$ more explicitly as random variables D and Z_i. Take as the inner product $\langle \mathbf{d}, \mathbf{z}\rangle = \mathbb{E}[DZ]$. This is known to be a Hilbert space. As above, let $G \subseteq H$ be a subspace spanned by N *orthonormal basis vectors,* that is, $G = \text{span}\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$. Explicitly write what this means for the random variables $Z_1, Z_2, \ldots, Z_N$.

(iii) (3 pts) Take your general result from Part (i) and write it now for the special Hilbert space from Part (ii), that is, in terms of standard random variable notation:

$$\min_{\hat{D}\in\text{span}\{\ldots\}} \ldots = \ldots \tag{28}$$

(iv) (3 pts) In class, we have carefully studied estimation problems with respect to the mean-squared error (MSE) criterion, identifying the MMSE estimator and the LMMSE estimator as well as their respective performance. See Section 6.2 of the Lecture Notes. Precisely explain the connection between your result from Part (iii) above to MSE estimation! Feel free to refer to formulas from the lecture notes simply by their equation number.

Solution 3. We take up the items in turn:

(i) As we have seen in class, Theorem 9.3 of the lecture notes, in the special case where we have an orthonormal basis of G, the minimizer takes a very simple shape, namely, it is given by $\hat{\mathbf{d}} = \sum_{i=1}^{N} \langle \mathbf{d}, \mathbf{z}_i\rangle \mathbf{z}_i$. One way to proceed is

$$\begin{aligned} \|\mathbf{d} - \hat{\mathbf{d}}\|^2 &= \langle \mathbf{d} - \hat{\mathbf{d}}, \mathbf{d} - \hat{\mathbf{d}}\rangle && (29)\\ &= \langle \mathbf{d}, \mathbf{d} - \hat{\mathbf{d}}\rangle - \langle \hat{\mathbf{d}}, \mathbf{d} - \hat{\mathbf{d}}\rangle && (30)\\ &= \langle \mathbf{d}, \mathbf{d} - \hat{\mathbf{d}}\rangle, && (31) \end{aligned}$$

where we have used the orthogonality principle, Theorem 9.1, to drop a term. This nicely simplifies our manipulations. Namely,

$$\|\mathbf{d}-\hat{\mathbf{d}}\|^2 = \langle \mathbf{d},\mathbf{d}\rangle - \langle \mathbf{d},\hat{\mathbf{d}}\rangle \tag{32}$$

$$= \|\mathbf{d}\|^2 - \langle \mathbf{d}, \sum_{i=1}^{N} \langle \mathbf{d},\mathbf{z}_i\rangle \mathbf{z}_i\rangle \tag{33}$$

$$= \|\mathbf{d}\|^2 - \sum_{i=1}^{N} \langle \mathbf{d},\mathbf{z}_i\rangle^* \langle \mathbf{d},\mathbf{z}_i\rangle \tag{34}$$

$$= \|\mathbf{d}\|^2 - \sum_{i=1}^{N} |\langle \mathbf{d},\mathbf{z}_i\rangle|^2. \tag{35}$$

Hence, we have found the result

$$\min_{\hat{\mathbf{d}}\in G} \|\mathbf{d}-\hat{\mathbf{d}}\|^2 = \|\mathbf{d}\|^2 - \sum_{i=1}^{N} |\langle \mathbf{d},\mathbf{z}_i\rangle|^2. \tag{36}$$

Alternatively, the following slightly longer approach also works:

$$\|\mathbf{d}-\hat{\mathbf{d}}\|^2 = \langle \mathbf{d}-\hat{\mathbf{d}}, \mathbf{d}-\hat{\mathbf{d}}\rangle \tag{37}$$

$$= \left\langle \mathbf{d} - \sum_{i=1}^{N} \langle \mathbf{d},\mathbf{z}_i\rangle \mathbf{z}_i, \mathbf{d} - \sum_{j=1}^{N} \langle \mathbf{d},\mathbf{z}_j\rangle \mathbf{z}_j \right\rangle \tag{38}$$

and then work through several steps of breaking the inner product, just like the steps we did together in class. This ultimately leads exactly to the same formula.

(ii) The first observation is that the abstract vectors of our general Hilbert space are random variables in the special Hilbert space under consideration here. So, we write $G = \text{span}\{Z_1, Z_2, \ldots, Z_N\}$, where Z_i are zero-mean finite variance random variables. Moreover, we know that this is an orthonormal basis. This means that $\langle \mathbf{z}_i, \mathbf{z}_i\rangle = 1$, or in our more explicit notation for the special Hilbert space at hand, $\mathbb{E}[Z_i^2] = 1$. It also means that for $i \neq j$, we have $\langle \mathbf{z}_i, \mathbf{z}_j\rangle = 0$, or in our more explicit notation for the special Hilbert space at hand, $\mathbb{E}[Z_i Z_j] = 0$.

(iii) Now, in Part (i), we found

$$\min_{\hat{\mathbf{d}}\in G} \|\mathbf{d}-\hat{\mathbf{d}}\|^2 = \|\mathbf{d}\|^2 - \sum_{i=1}^{N} |\langle \mathbf{d},\mathbf{z}_i\rangle|^2. \tag{39}$$

Specializing to the Hilbert space at hand, we thus write

$$\|\mathbf{d}-\hat{\mathbf{d}}\|^2 = \langle \mathbf{d}-\hat{\mathbf{d}}, \mathbf{d}-\hat{\mathbf{d}}\rangle = \mathbb{E}[(D-\hat{D})(D-\hat{D})] = \mathbb{E}[(D-\hat{D})^2] \tag{40}$$

and

$$\|\mathbf{d}\|^2 - \sum_{i=1}^{N} |\langle \mathbf{d}, \mathbf{z}_i \rangle|^2 = \langle \mathbf{d}, \mathbf{d} \rangle - \sum_{i=1}^{N} |\langle \mathbf{d}, \mathbf{z}_i \rangle|^2 \tag{41}$$

$$= \mathbb{E}[D^2] - \sum_{i=1}^{N} (\mathbb{E}[DZ_i])^2. \tag{42}$$

Combining, this gives us

$$\min_{\hat{D} \in \text{span}\{Z_1, Z_2, \ldots, Z_N\}} \mathbb{E}[(D - \hat{D})^2] = \mathbb{E}[D^2] - \sum_{i=1}^{N} (\mathbb{E}[DZ_i])^2. \tag{43}$$

(iv) Part (iii) is precisely the Linear MMSE estimation problem of Section 6.2.2. In there, the estimate $\hat{D}$ is formed based on the observed random vector $\mathbf{X}$. Here, we denote this random vector by $\mathbf{Z}$. As observed above, the components of this vector are orthonormal, meaning that $\mathbb{E}[Z_i Z_j] = 0$ and $\mathbb{E}[Z_i^2] = 1$. In this special case, the matrix $R_{\mathbf{Z}}$ in Equation (6.13) of the lecture notes is simply the identity matrix. Therefore, the optimal coefficients in Equation (6.15) are given by $\mathbf{w} = \mathbb{E}[D\mathbf{Z}]$. Therefore, the incurred error in Equation (6.20) is given by $\mathbb{E}[D^2] - \mathbb{E}[D\mathbf{Z}]^H \mathbb{E}[D\mathbf{Z}] = \mathbb{E}[D^2] - \sum_{i=1}^{N} (\mathbb{E}[DZ_i])^2$. This is precisely the formula we found in Part (iii).

Problem 4 (The Maximum Likelihood Estimator and I-projections – 10 pts)**.** Let X^n be i.i.d. random variables taking values in a finite set $\mathcal{X}$ of size k. Recall that when restricted to distributions from a set $\mathcal{P}$, the maximum likelihood estimator $\hat{p}_{\text{MLE}} : \mathcal{X}^n \to \mathcal{P}$ is given as

$$P^*_{\text{MLE}} = \hat{p}_{\text{MLE}}(x^n) = \arg\max_{P \in \mathcal{P}} P\left(X^n = x^n\right).$$

Note that the two parts of this problem are independent of each other.

(i) (4 pts) Show that the maximum likelihood estimate P^*_{MLE} is equal to the I-projection

$$P^* := \arg\min_{P \in \mathcal{P}} D\left(\hat{P} \middle\| P\right)$$

where $\hat{P}$ denotes the empirical distribution of x^n.

(ii) (6 pts) Let P_0 be a distribution over $\mathcal{X}$ and let $\mathcal{P}$ be defined as the minimal exponential family

$$\mathcal{P} := \left\{P : P(x) = P_0(x)e^{-\langle\theta,\phi(x)\rangle - A(\theta)}, \theta \in \Theta\right\}$$

for some open parameter space $\Theta \subseteq \mathbb{R}^d$, sufficient statistic ϕ, and normalization function A. Then show that the maximum likelihood estimate P^*_{MLE} is the I-projection P^* of P_0 onto the linear family

$$\mathcal{L}(x^n) := \left\{P : E_P\left[\phi(X)\right] = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right\}.$$

In other words, show that

$$\arg\min_{P \in \mathcal{L}(x^n)} D\left(P_0 \| P\right) = \arg\max_{Q \in \mathcal{P}} Q(X^n = x^n).$$

Solution 4. 1. Since the logarithm is an increasing function, we can equivalently maximize the log likelihood of the sequence x^n as follows:

$$\begin{aligned}
\arg\max_{P\in\mathcal{P}} P\left(X^n = x^n\right) &= \arg\max_{P\in\mathcal{P}} \log P\left(X^n = x^n\right) = \arg\max_{P\in\mathcal{P}} \log \prod_{i=1}^{n} P(x_i) \\
&= \arg\max_{P\in\mathcal{P}} \sum_{i=1}^{n} \log P(x_i) \\
&\overset{(a)}{=} \arg\max_{P\in\mathcal{P}} \sum_{x\in\mathcal{X}} n\hat{P}(x) \log P(x) \\
&\overset{(b)}{=} \arg\min_{P\in\mathcal{P}} \sum_{x\in\mathcal{X}} n\left(\hat{P}(x)\log\hat{P}(x) - \hat{P}(x)\log P(x)\right) \\
&= \arg\min_{P\in\mathcal{P}} nD\left(\hat{P}\middle\|P\right) = \arg\min_{P\in\mathcal{P}} D\left(\hat{P}\middle\|P\right),
\end{aligned}$$

where (a) is true because the term $\log P(x)$ is added exactly $n\hat{P}(x)$ times in the summation for each $x \in \mathcal{X}$, and (b) follows from the fact that adding the constant term $n\sum_{x\in\mathcal{X}} \hat{P}(x)\log\hat{P}(x)$ to the objective does not change the minimizing argument.

2. From section 8.5 of the notes, we know that when $\mathcal{P}$ is the exponential family with base density h and sufficient statistic ϕ, the MLE also lies in the linear family $\mathcal{L}$. Also, from Theorem 8.3 in the notes, we know that the I-projection of P onto the linear family $\mathcal{L}$ lies precisely the intersection of the exponential family with the linear family. Invoking Theorem 8.4 for the uniqueness of θ, since the sample mean lies in the set of feasible means $\mathcal{M} := [\min_x \phi(x), \max_x \phi(x)]$ with probability 1 *(why?)*, the proof is complete.

Alternate method for showing that $P^* \in \mathcal{P} \cap \mathcal{L}$:

Consider the problem of finding the I-projection of P_0 onto $\mathcal{L}$:

$$P^* = \arg\min_{P\in\mathcal{L}} D\left(P\|P_0\right)$$

Using Lagrange multipliers once again, write the Lagrangian function $L(\mu_1, \mu_2)$ as

$$\sum_{x\in\mathcal{X}} \left(P(x)\log P(x) - P(x)\log P_0(x) + \mu_1 P(x)\left(\phi(x) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right)\right) + \mu_2\left(\sum_{x\in\mathcal{X}} P(x) - 1\right)$$

Taking partial derivatives w.r.t. $P(x)$and equating to zero, we get

$$\begin{aligned}
\frac{\partial}{\partial P(x)} L(\mu) &= 1 + P(x) - \log P_0(x) + \mu_1\phi(x) - \mu_1\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) + \mu_2 = 0 \\
\implies P(x) &= P_0(x)\exp\left(-\mu_1\phi(x) - \left(\mu_2 + 1 - {}^{\mu_1}\!/_n \sum_{i=1}^{n}\phi(x_i)\right)\right).
\end{aligned}$$

Writing $-\mu_1$ as θ and $\mu_2 + 1 - {}^{\mu_1}\!/_n \sum_{i=1}^{n}\phi(x_i)$ as $A(\theta)$, we find that $P^* \in \mathcal{P}$.